Summary: Via 5 regression methods, this paper evaluates the effect of air pollution **nox** on 506 districts' median property values **mv** controlling for effects of 11 other dataset features.

Research design: Focus is on visual and numeric bivariate relationships with **mv** and **nox**, and then on the results of multiple linear, ridge, lasso and ElasticNet regression methods revealed by cross validation, learning curves, and by varying alphas, detailed in the Python code section.

Table 1:
Number and % of columns with missing values in the data set:

	Train na	Train %na
crim	0	0
zn	0	0
indus	0	0
chas	0	0
nox	0	0
rooms	0	0
age	0	0
dis	0	0
rad	0	0
tax	0	0
ptratio	0	0
lstat	0	0
mv	0	0

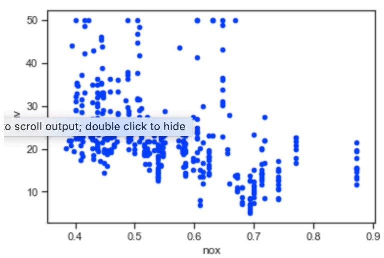
Table 2a: Modes

	mode	freq
crim	0.02	2.0
zn	0.00	296.0
indus	18.10	104.0
chas	0.00	377.0
nox	0.54	21.0
rooms	5.71	3.0
age	100.00	36.0
dis	3.50	5.0
rad	24.00	104.0
tax	666.00	104.0
ptratio	20.20	110.0
lstat	8.05	3.0
mv	50.00	13.0

Table 2b:
nox and mv modes and counts

	nox mode	count	mv mode	count
0	0.538	21.0	50.0	13.0
1	0.437	15.0	19.4	6.0
2	0.605	13.0	21.7	6.0
3	0.713	12.0	20.6	5.0
4	0.740	12.0	19.6	5.0
5	0.871	12.0	20.1	5.0
6	0.624	12.0	20.0	5.0
7	0.647	10.0	22.0	5.0
8	0.700	10.0	23.1	5.0
9	0.693	10.0	25.0	5.0
10	0.489	10.0	22.6	5.0
11	0.547	9.0	19.9	4.0

Figure 0: Repeated values in mv and nox



Data Exploration / Preparation: (1) **Table 1** above shows a dataset has no missing values. (2) There are 506 samples, 13 features and 1 response. (3) Preparation drops the only categorical feature and splits the rows by **nox** proportions into training and test sets. (4) Repeated values that are found in (a) **figure 1**'s *pair-wise plots* on the last page, (b) **figure 0's** plot of mv vs nox above and in (c) **table 2a & b** above containing variable modes and their counts, may arise from missing value imputations, and may confound our machine learning. (5) Excepting the response variable, no univariate plot in **figure 1** appears normal. (6) **Table 3** on the last page computes correlations amongst the 13 columns containing 10 ratio level data, 1 binary=**chas**, 1 categorical=**zn**, and 1 ordinal=**rad**. **Table 4's** presents a data dictionary and salient relationships. Most pertinent are these correlations: **mv-nox** 43%, **nox-ptratio** 19%, **nox-rooms**

-33%, **mv-ptratio** 51%, **mv-rooms** 72%. While **mv-lstat**=75%, **nox-lstat** = 59% suggests that much of **lstat**'s effect on **mv** may already be collected in **nox**.

```
Table 4:
          Description         Type  visual mv  corr mv
crim       crime rate        ratio
zn        %zone 4 lots  categorical
indus      %bus indus        ratio
chas       on charles       binary
nox           air pol        ratio
rooms      rooms/home        ratio  Pos Assoc      0.7
age          % < 1940        ratio
dis        dist 2 empl       ratio
rad        hway access     ordinal
tax          tax rate        ratio
ptratio    stud/teach        ratio
lstat     %low ecosoc        ratio  Neg Assoc    -0.74
mv         median val        ratio


Notes on corr matrix:
 mv: rooms +72, lstat -74, ptratio -51, nox -43
 nox: indus +76, age +73, dis -77, lstat +59, tax +67, rad +61
 nox: ptratio only +19, rooms -33
 lstat: rooms -63, indus +59, age +60
 rooms: for zn!=0 rooms +93 w mv
 dis: nox -77 age -75, zn +66, indus -71
 age: nox +73, dis -75, indus +65, lstat +60
 indus: tax +72, ptratio
 crim: non linear w mv

Possible key relationships:
mv ~ (nox -43) + (nox 19 ptratio -51 mv) + (nox -33 rooms 72 mv)
```

<u>Python code</u>: (1) The data set is split into 80% train / 20% test stratified by the key feature **nox**. Here's why. The 1st of 4 equal width **nox** bins in **figure 2**'s histogram is nearly distinct from the remaining bins in **figure 3**'s box plot of **mv**; getting their proportions correct may affect cross validation results. **Table 4** demonstrates how the 4th bin is very underrepresented in random splitting, and so stratified samples are collected for the analyses which follow.

Figure 2: Histogram of nox categories

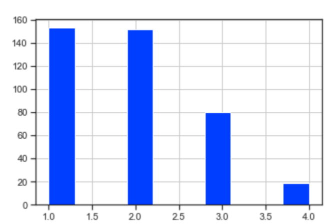


Figure 3: Box plot of mv binned by nox categories

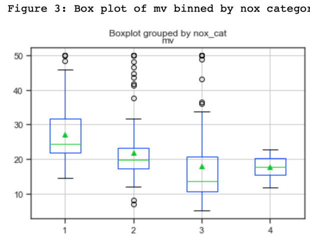


Table 4: Differing proportions for random and stratified sampling:

	Overall	Stratified	Random	Rand. %error	Strat. %error
1	0.379447	0.382353	0.392157	3.349673	0.765931
2	0.375494	0.372549	0.392157	4.437564	-0.784314
3	0.197628	0.196078	0.196078	-0.784314	-0.784314
4	0.047431	0.049020	0.019608	-58.660131	3.349673

(2) A single pipeline transformation standardizes all variables via sklearn's StandardScaler. Tibshirani, (1995) suggests replacing **zn's** categories with dummy variables before standardizing (p. 394), but the variable **rooms** is already 93% correlated with **mv** for the 25 **non-zero values**

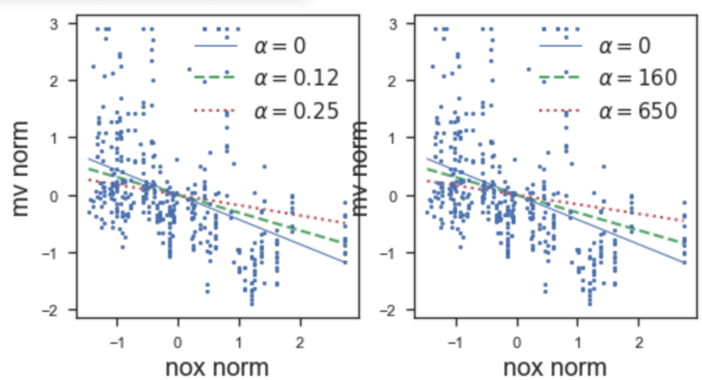
of zn. With more time, a single binary for **zn non zero** might be added but **rooms** may already reflect **zn's** variances and **room** carries 71% correlation with **mv (**versus **mv-zn** = 37.5%. (3) **Table 5** presents a 10 fold cross validation which compares RMSE and model parameters

Table 5:
Average results from 10-fold cross-validation
in standardized units (mean 0, standard deviation 1)

Method	Root mean-squared error
Linear_Regression	0.521499
Ridge_Regression	0.520371
Lasso_Regression	0.572645
ElasticNet	0.560228

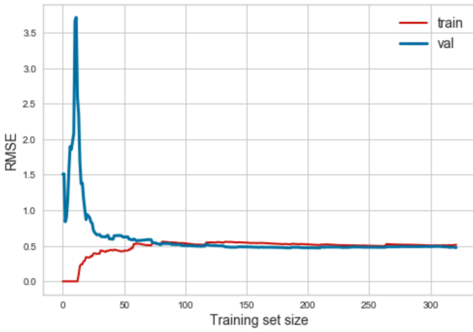
for (a) multiple linear regression (MLR), (b) ridge regression (RR) with alpha = 10, (c) Lasso with alpha = 0.1 and (d) ElasticNet (EN) with alpha = 0.1 evenly split between Lasso-RR. RMSE results favor MLR and a mild form of RR. (4) **Figure 4** runs Lasso (LHS) and RR (RHS) regressions for **mv vs nox** with alphas set to achieve

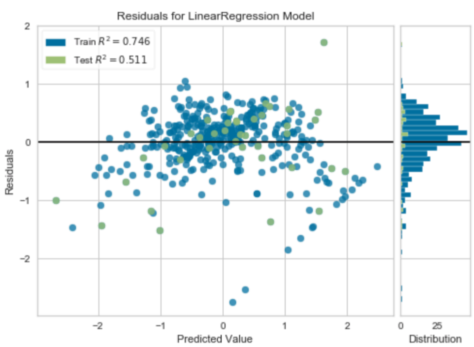
Figure 4: Same regressions from
different alphas in Lasso LHS and Ridge-RHS
click to scroll output; double click to hide



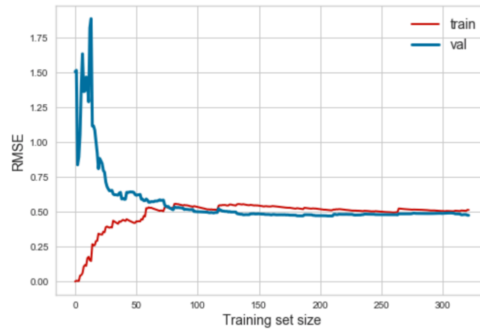
similar results. The alpha magnitude differences accrue to L1 employed by Lasso. L1 has a stronger effect than L2 in RR thus a lower alpha can be used for Lasso. (5) Learning curves and residual plots for each regression method are presented in **figures 5a-d**. In all cases, as samples are added, training RMSE is higher than validating RMSE. These converge for MLR and RR but not Lasso and EN. These results are not cross validated; it is possibly due to sampling methods that favor the validation set, but a more logical cause is that regularization costs are included in the training but not the test fit RMSE.

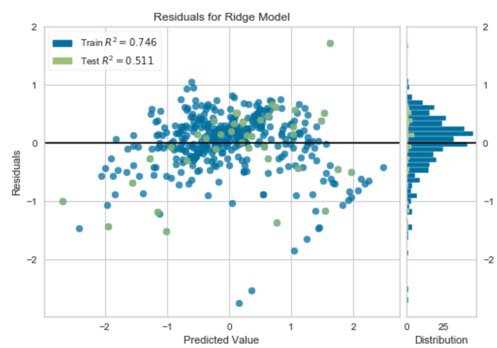
```
Scikit Learn method: Linear_Regression
Fitted regression coefficients:
 [-0.104  0.137  0.013  0.092 -0.223  0.308 -0.014 -0.35   0.261 -0.248
 -0.201 -0.388]
Root mean-squared error: 0.5038331339844982
```



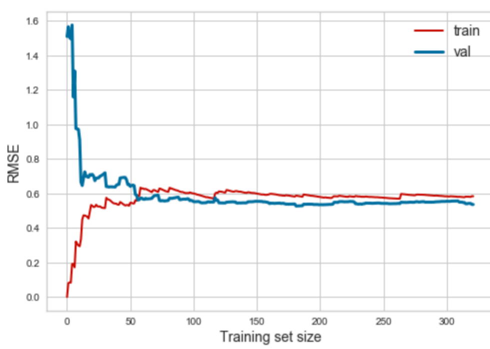


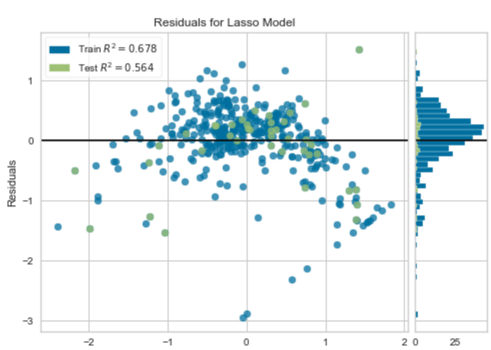

```
Scikit Learn method: Ridge_Regression
Fitted regression coefficients:
 [-0.104  0.137  0.012  0.092 -0.222  0.309 -0.014 -0.35   0.26  -0.247
 -0.2   -0.388]
Root mean-squared error: 0.5038333286021991
```



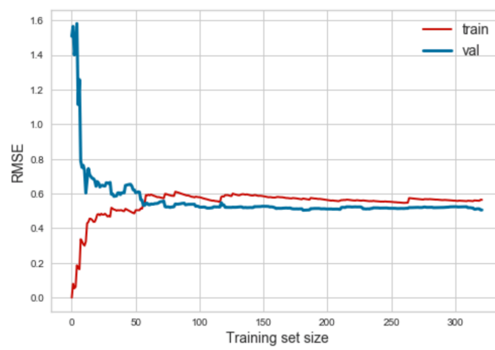


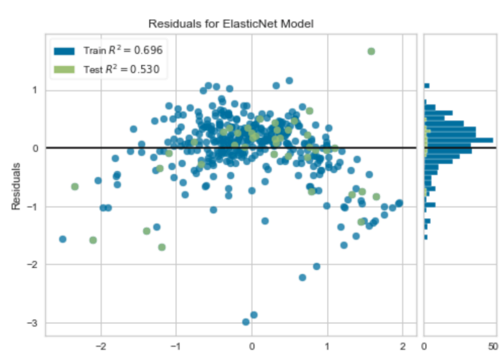

```
Scikit Learn method: Lasso_Regression
Fitted regression coefficients:
 [-0.     0.    -0.     0.014 -0.     0.323 -0.    -0.    -0.    -0.
 -0.148 -0.38 ]
Root mean-squared error: 0.5672493128388454
```






```
Scikit Learn method: ElasticNet
Fitted regression coefficients:
 [-0.019  0.    -0.     0.058 -0.     0.343 -0.    -0.005 -0.    -0.034
 -0.164 -0.366]
Root mean-squared error: 0.5509688624562946
```





Management recommendations.

Which model? While MLR and RR present the lowest RMSE, the flexibility of EN provides significant power to scale into a more sophisticated model: with L1 = 0, EN is the same as RR and with alpha = zero, EN will have the same results as MLR. Go with EN for flexibility; set it to MLR equivalents to start with.

Partial coefficients As to the question of air pollution impact has on median values, the answer depends upon the model. With Lasso and EN, the nox coefficient is zero as L1 regularization favors 4 features in Lasso and 7 features in EN, and zeros out the remaining 12 features regressed. With MLR and RR, our standardized regression formula above provides:

normalized mv =

-0.104*crim

+0.137*zn

+0.012*indus

+0.092*chas

-0.222*nox

+0.309*rooms

-0.014*age

-0.35*dis

+0.26*rad

-0.247*tax

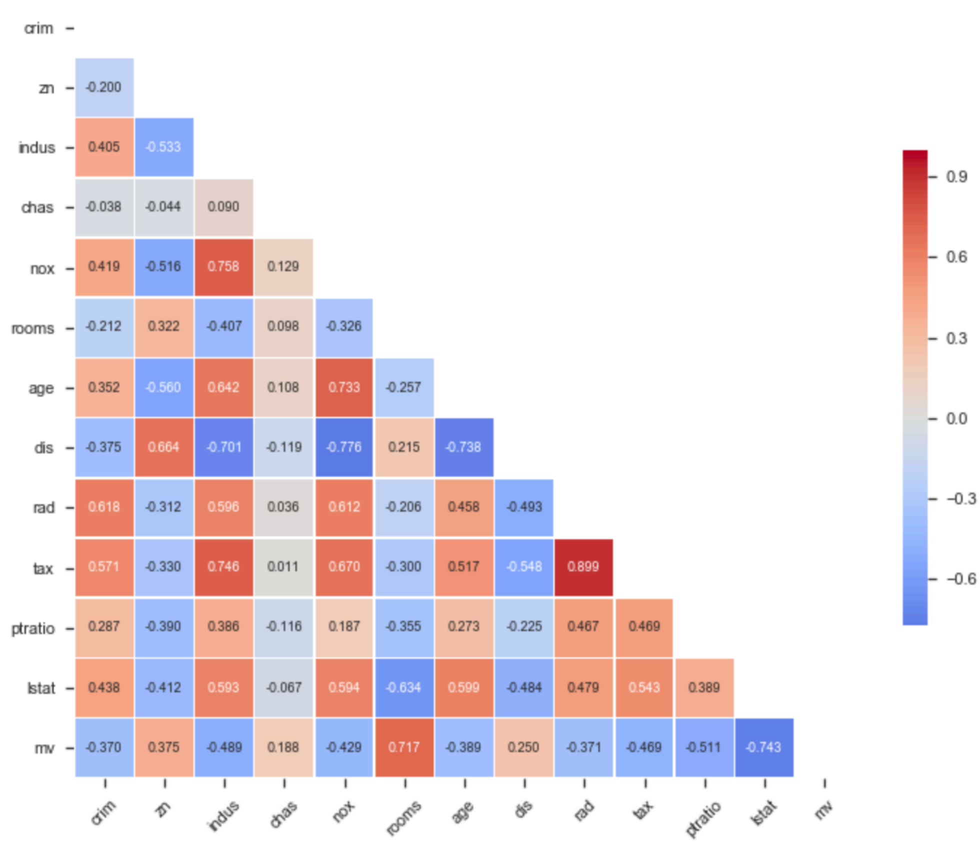
-0.2*ptratio

-0.388*lstat

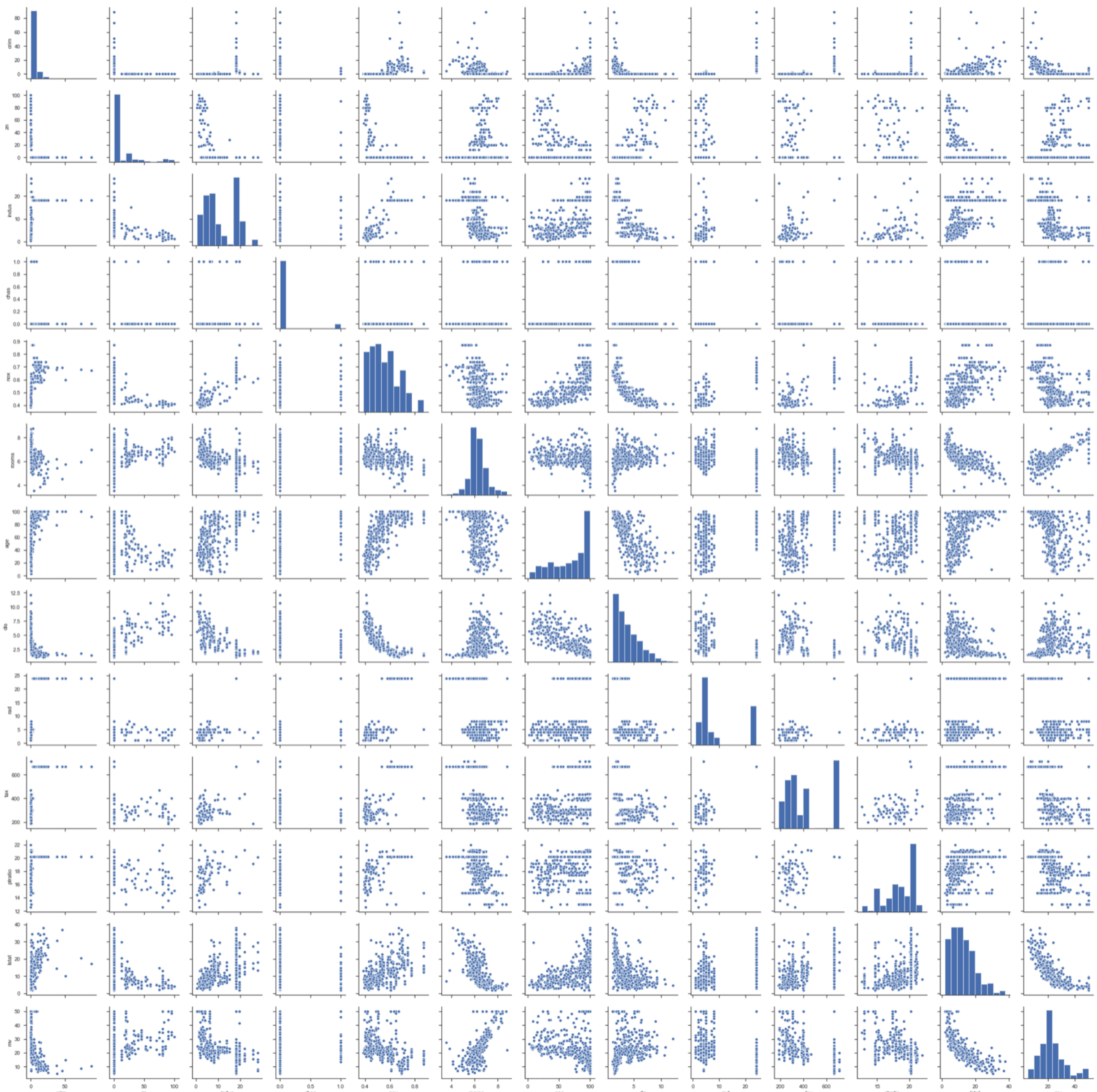
To obtain the de normalized coefficient, -18.08, simply divide the normalized coefficient for nox, -0.222 above, by the scale_ variable for X for nox (1.156e-01) and multiply by the scale_ variable for y for mv (9.414e+00). These scale_ values can also be obtained as simple standard deviations of the nox and mv data columns, adjusted for DOF. Mean values do not enter into the denormalization since the mean of the standardized variables is zero. -18.08 represents the impact that air pollution has on the median value of property, controlling for other variables movement in the model: As air pollution rises by 1, and all other variables are held constant, median property value declines by 18.08.

Table 3: Correlations

<Figure size 576x396 with 0 Axes>



```
Figure 1: Pairwise plots and univariate histograms:
```



References:

TIBSHIRANI, R. (1997), THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL. Statist. Med., 16: 385-395. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3